SAS® GLOBALFORUM 2015

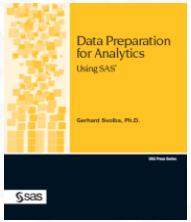
SAS1440-2015

The Journey Is Yours

Want an Early Picture of the Data Quality Status of Your Analysis Data? SAS® Visual Analytics Shows You How

Gerhard Svolba, SAS Institute Inc. - Austria

Dallas, TX – April 29th, 2015



SAS GLOBALFORUM

From this presentation you can expect

- The statisticians' view on data quality
- Meeting my Aunt Susanne
- SAS macros and SAS programs to see your data from the bird's eye view
- Live SAS®Visual Analytics software demo to profile your analysis data

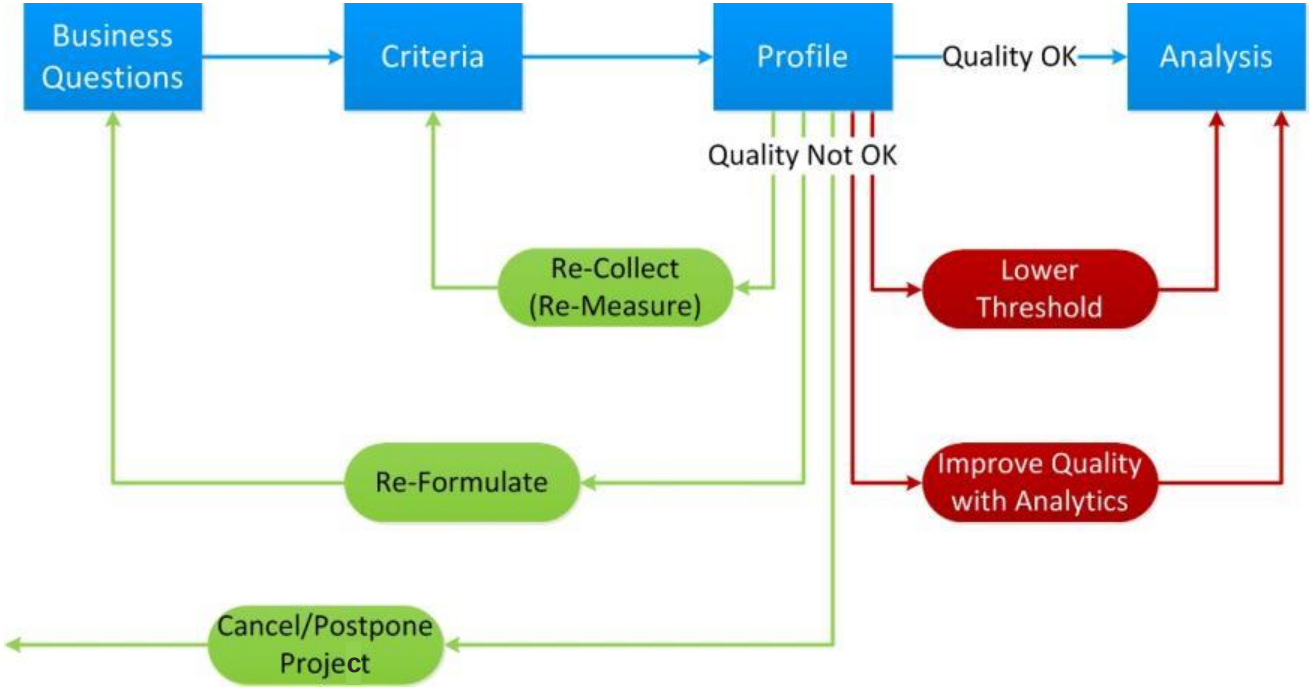
Idea and rationale of analytical data quality profiling

- To gain a picture of the usability of the data for analytical methods early in the analytical lifecycle
  - Time Savings
  - Cost Savings
  - Creation of an „Analytic Awareness“
    → Analytic Data Quality Monitoring should already start there, where Analytical Data Marts are created.

Data quality is an analytic topic!

- Analysis data might be fine from a technical perspective. But analytic requirements go beyond these items.
  - Historic data and historic snapshots of the data are not the same
  - Correlation between variables can help. But might cause a headache (substitution effects vs. multicollinearity)
  - Missing values: the number of usable observations for the analysis reduces quickly
  - Systematic pattern in missing values and outliers
  - Distribution of variables

Bad data quality has consequences!



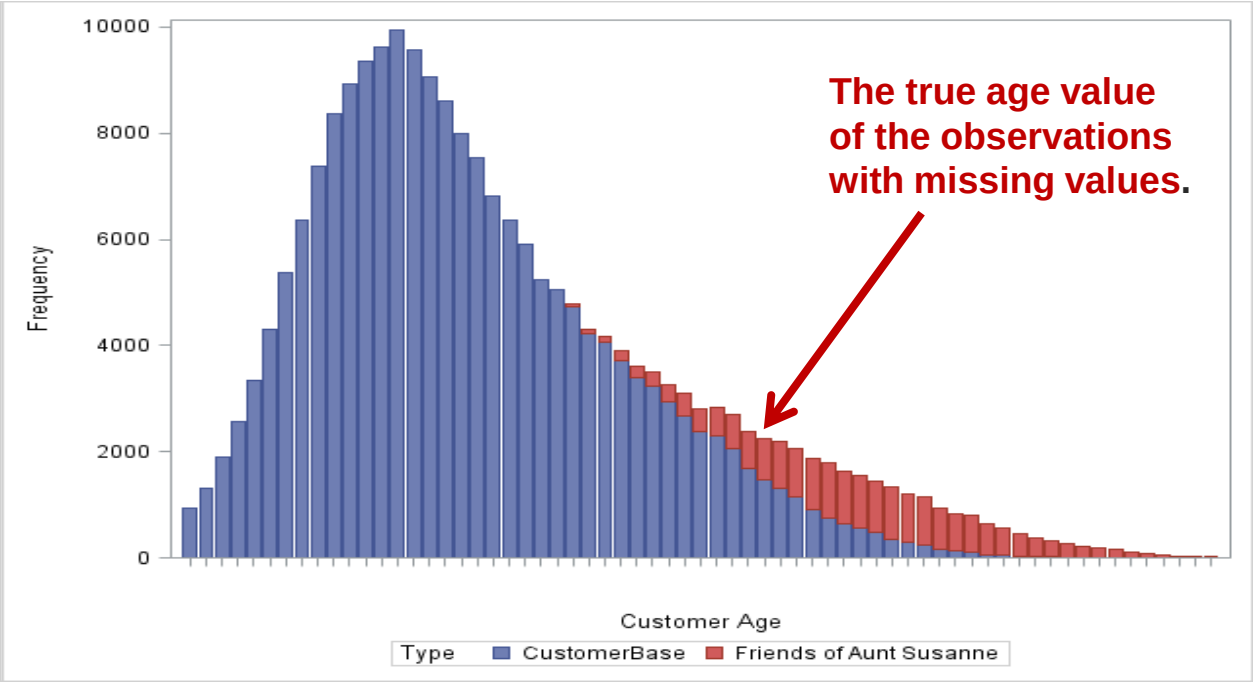
Cost
Time, Delays
No Results

Trust
Risk of wrong decisions
Insignificance

Why my Aunt Susanne gives analysts a hard time

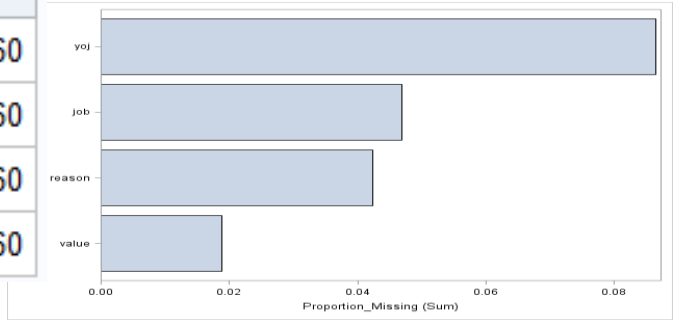
- She got her phone in the mid 1960s.
- Customers‘ „Date of birth“ was of no interest at that time.
- Since the mid 1990s it is mandatory to provide the date of birth on a new contract.
- She never changed her contract type or answered any customer questionnaires.
- She is not the only one with this „data history“.

What does her phone provider see, when he looks at the customer age variable



Typically missing values are analyzed in a univariate way

Variable	Frequency_Missing	Proportion_Missing	N
YOJ	515	8.64%	5960
JOB	279	4.68%	5960
REASON	252	4.23%	5960
VALUE	112	1.88%	5960



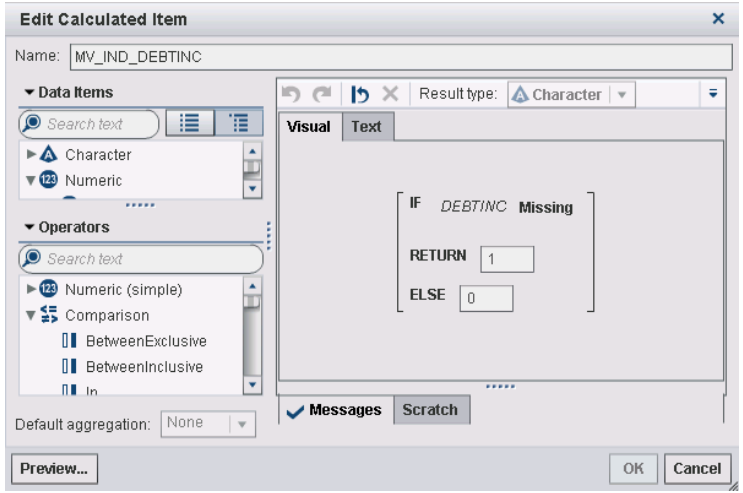
- How many of your variables are infected by the “missing value disease”?
- Not: How many “Full-Records“ do you have?
- Not: Is there a pattern in the structure of missing data?

How can you detect such situations?

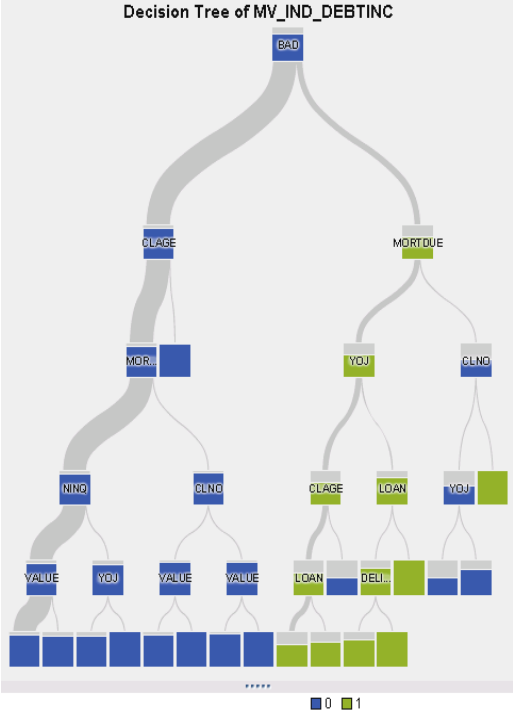
- Simple frequencies per variable do not help.
- Create an indicator variable „Missing YES/NO“ and compare the distribution of other variables like customer start date, product portfolio, …
- Business and process knowledge about the company is key!
- Define imputation rules based on expert knowledge.

Using a predictive model to explain the "missing yes/no" indicator

Step 1: Create a derived variable for Missing Yes/No

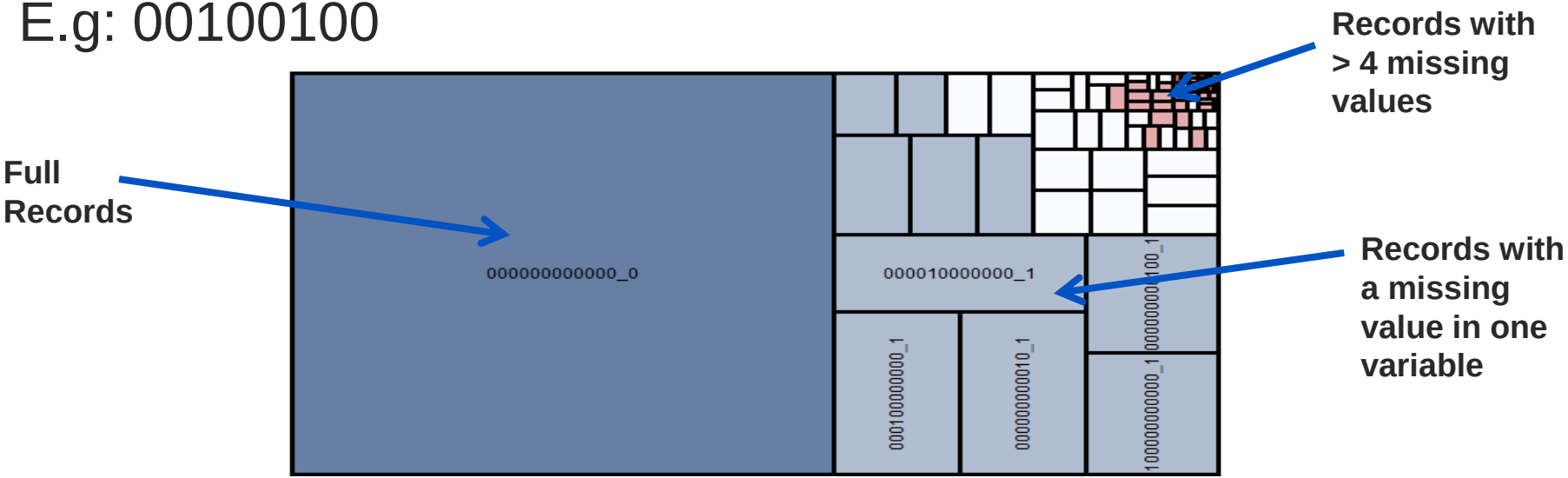


Step 2. Train a decision tree



Profiling the pattern of missing values with the %MV_PROFILNG macro

- Concatenate each “Missing-Value” Indicator to a string. E.g: 00100100

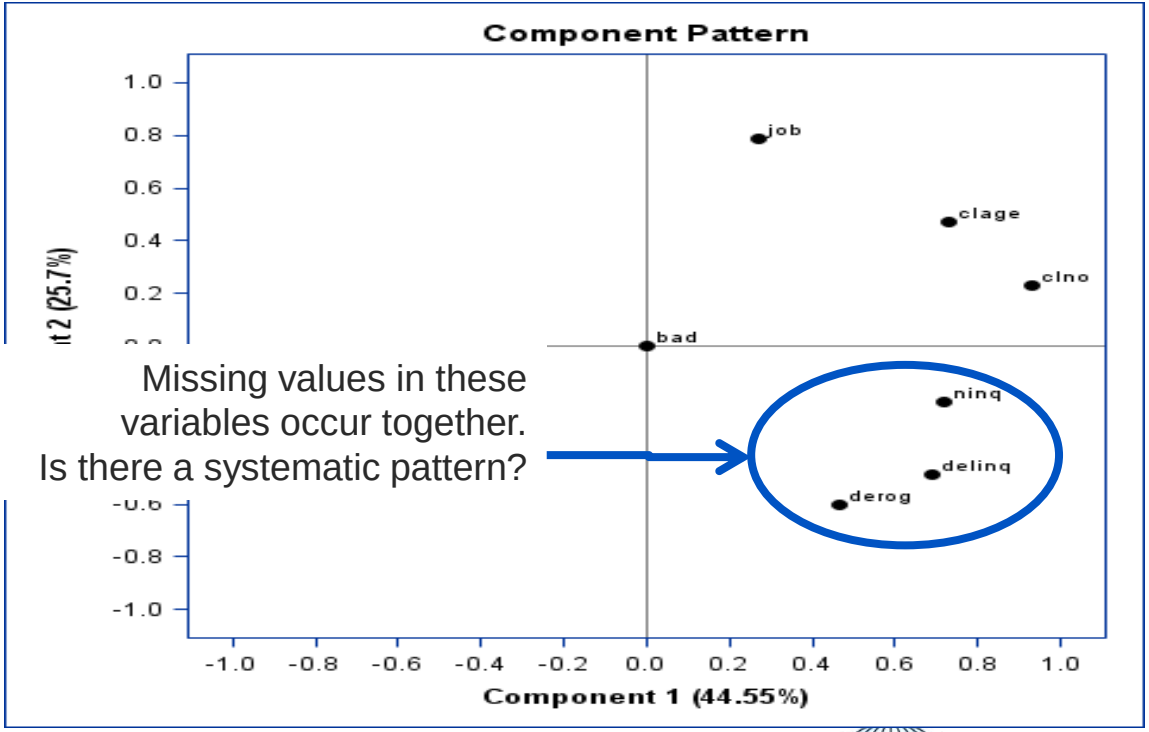


- Macros can be downloaded from sascommunity.org!

Multivariate analysis uncovers systematic patterns %MV_PROFILING macro

Principal Components

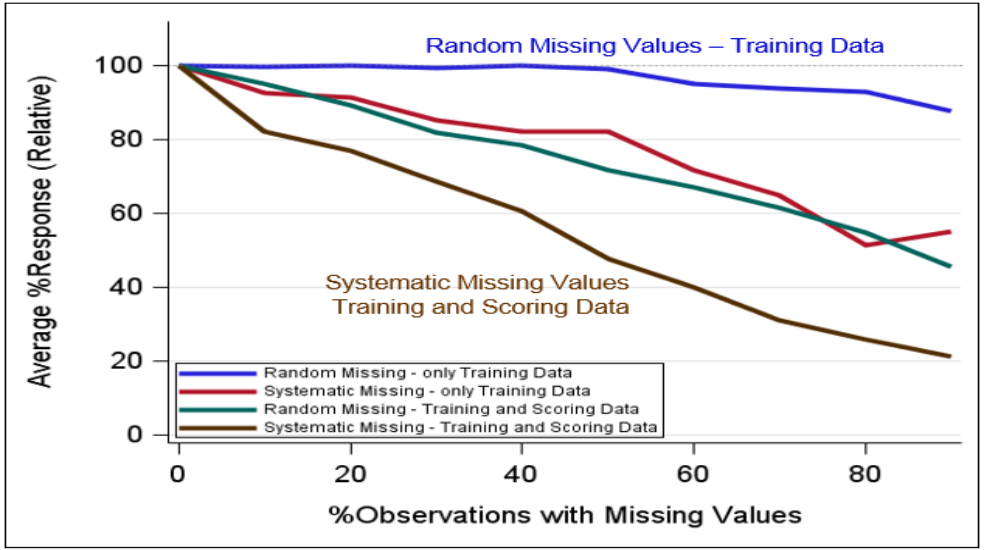
```
%MV_PROFILING(
    data=em.hmeq,
    vars=_ALL_,
    ODS=YES,
    varclus=NO,
    princomp=YES,
    ncomp=2,
    order=ALPHA);
```



Missing values in these variables occur together. Is there a systematic pattern?

Results of simulation studies: Does quantity matter? Not always.

- Random and systematic missing values have been inserted into the training and scoring data.
- Mean has been used for missing imputation.
- Results show little decrease in model accuracy with increasing proportion of missing values.
- Results show heavy differences between random and systematic missing values.



No missing values does not necessarily mean complete data

PNR	date	amount
56	2004-02-01	48
56	2004-03-01	51
56	2004-04-01	42
56	2004-05-01	36
56	2004-06-01	6
56	2004-07-01	.
56	2004-08-01	48
56	2004-09-01	36
56	2004-10-01	66
56	2004-11-01	15
56	2004-12-01	33
58	2005-06-01	39
58	2005-07-01	63
58	2005-08-01	84
58	2005-09-01	18
58	2005-12-01	69
58	2006-03-01	0
58	2006-07-01	90
58	2006-10-01	57
58	2007-01-01	48

Existing Record
Value Missing

?

Missing Record
No Continuity

Replacing and interpolating missing values in longitudinal data with SAS procedures

Insert missing records | Replace with 0 | Replace with last known value | Replace with mean | Interpolate based on splines (PROC EXPAND)

	DATE	air_mv	air_mv_zero	air_mv_previous	air_mv_mean	air_expand
1	JAN49	112	112	112	112	112
2	FEB49	118	118	118	118	118
3	MAR49	132	132	132	132	132
4	APR49	129	129	129	129	129
5	MAY49	.	0	129	284.54385965	128.29783049
6	JUN49	135	135	135	135	135
7	JUL49	.	0	135	284.54385965	144.73734152
8	AUG49	148	148	148	148	148
9	SEP49	136	136	136	136	136
10	OCT49	119	119	119	119	119
11	NOV49	.	0	119	284.54385965	116.19900978
12	DEC49	118	118	118	118	118
13	JAN50	115	115	115	115	115
14	FEB50	126	126	126	126	126
15	MAR50	141	141	141	141	141

```
PROC TIMESERIES
      DATA = air_missing
      OUT = timeid_inserted;
  ID date INTERVAL=MONTH
          SETMISS=0;
  VAR qty;
  BY prod_id;
RUN;
```

Use PROC TIMESERIES and PROC EXPAND for these tasks!

Profiling the structure of missing values with the %PROFILE_TS_MV macro

TS_Profile_Chain	Frequency	Percent
111111111111111111111111111111111111111111111111111111_54_0	18	39.13
1111111111111111111111111111111111111111111111111111111111_60_0	17	36.96
0000001111111111111111111111111111111111111111111111111111111_60_0	5	10.87
1111110011111111110000111111111111111111111111111111111111000001_60_0	1	2.17
1111111111111111111111111000000000000011111111111111111111111_60_0	1	2.17
1111111111111111111111111100111111111111111111111111111111111111_60_0	1	2.17
111111111111111111111111111111111111111111111111111111_53_0	1	2.17
1111111111111111111111111111111X111111111111111XX1X1XX11111XXXX_60_10	1	2.17
1111XX1111111111111111111111111111111111111X1X1XXXX11111XX111_60_10	1	2.17

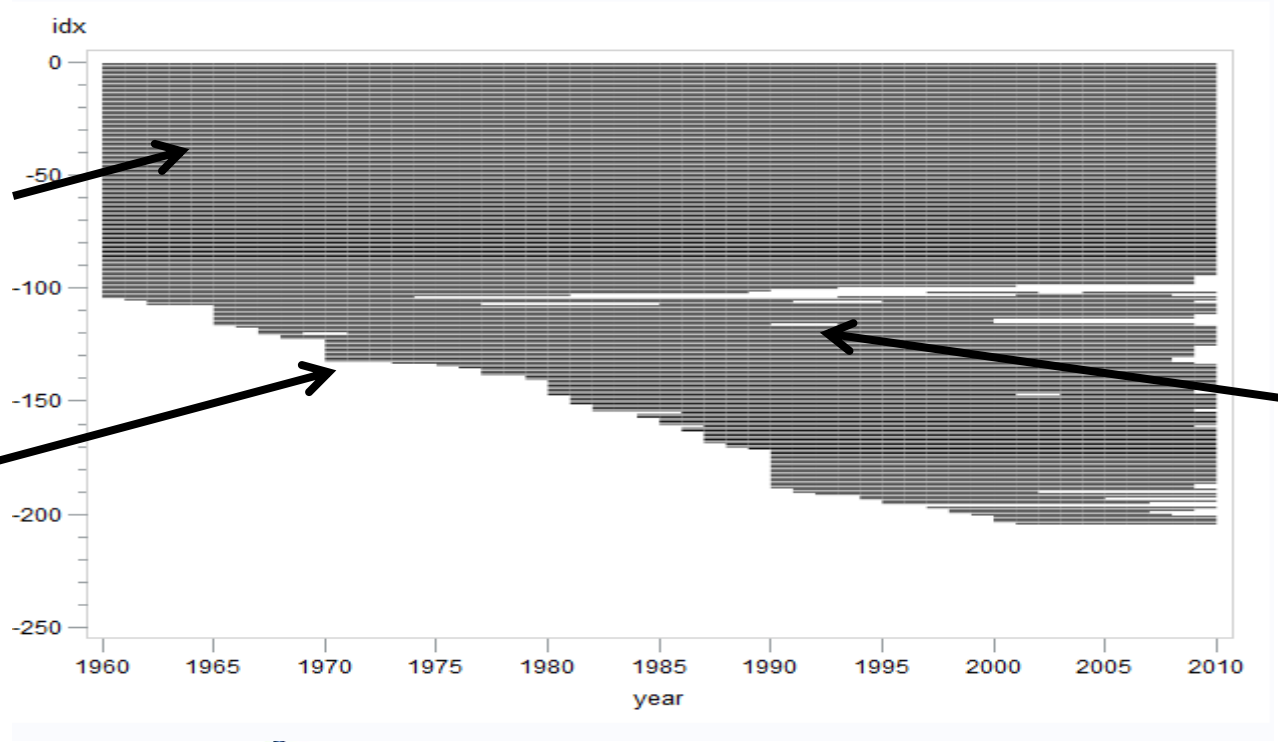
0 Value

Missing Value

Profiling the structure of missing values with the %PROFILE_TS_MV macro

Series with full length

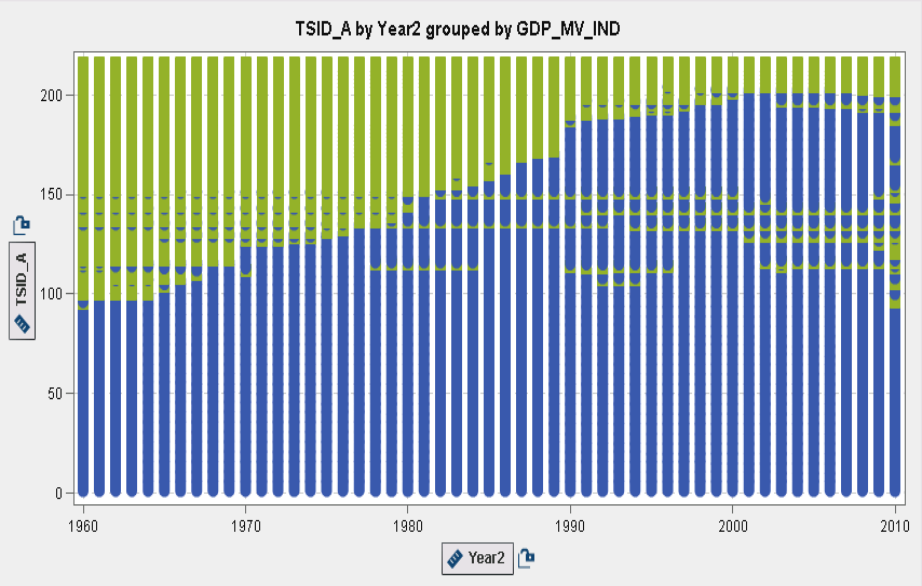
Series starting later



Series with embedded missings

```
%Profile_TS_MV(DATA=tmp.gdp2,ID=Country_name, DATE=Year, VALUE=gdp,
                MV=(.),W=1, NMAX_TS=300);
```

Using SAS® Visual Analytics to profile time series data



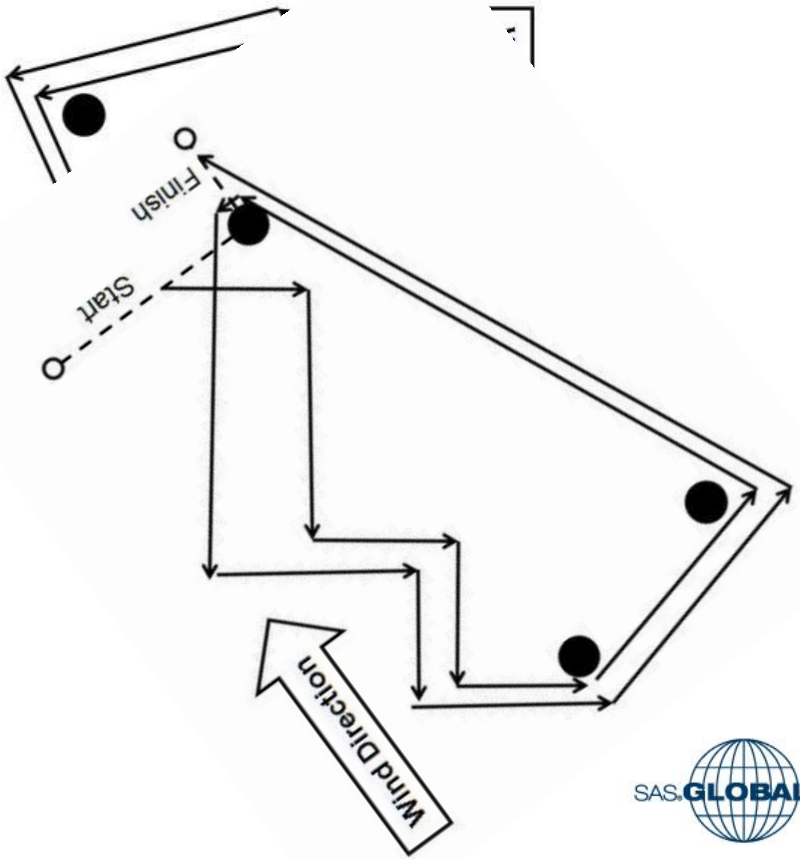
3 Steps to a bird's-eye view

1. Create a missing value indicator for GDP
2. Create a scatter plot with YEAR on the x-axis TSID on the y-axis
3. Add the missing value indicator to the plot

Using SAS Visual Analytics to find hidden data quality problems - Main benefits

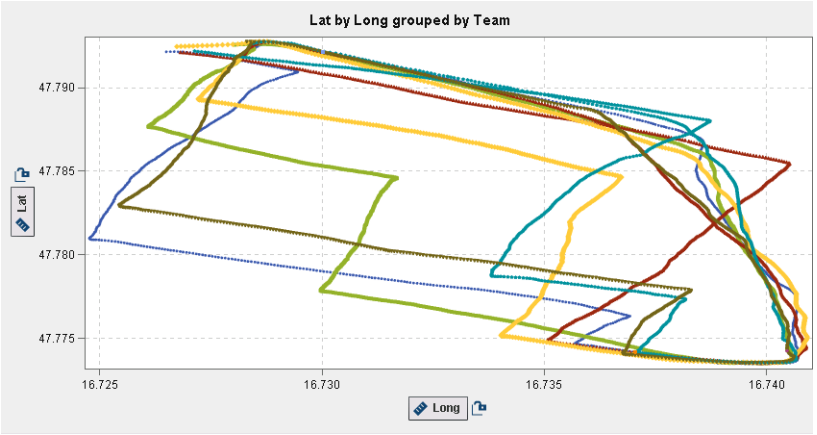
- You work directly on the data:
  - Filter, subset, and group your data while you explore it
- You are able to analyze value combinations from the business point of view:
  - Trace your findings directly to the source data.
- Two Examples:
  - Using decision trees to find the reason for data quality problems
  - Using interactive data analysis to uncover impossible value combinations

Layout of a sailing race with 3 buoys



Profiling GPS track point data

Displaying the race course by team



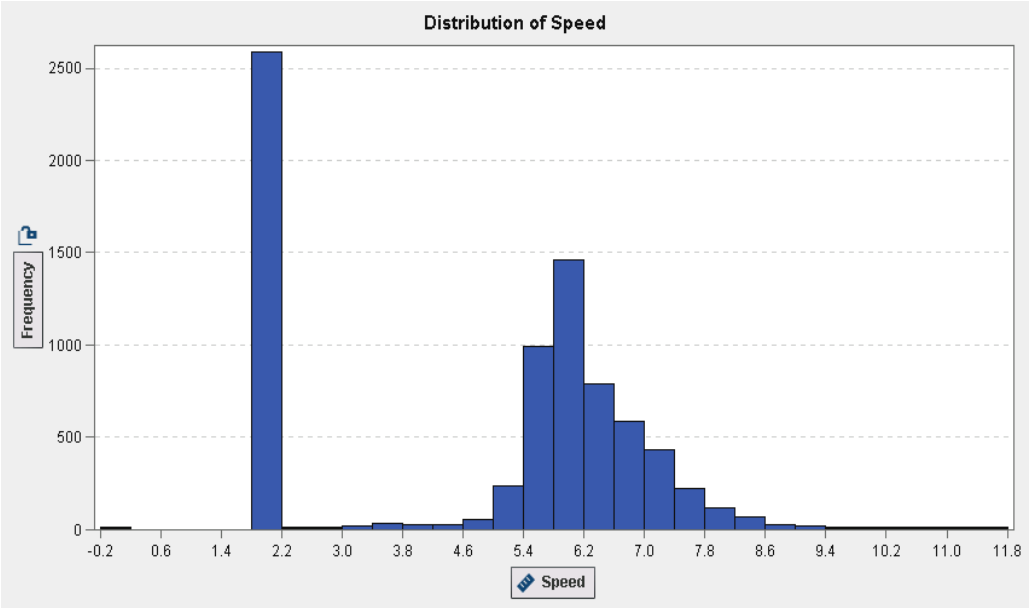
- courses exhibit a smooth, non-erratic behavior

Displaying the race course on a geo map

- points are located in the blue area (the lake)

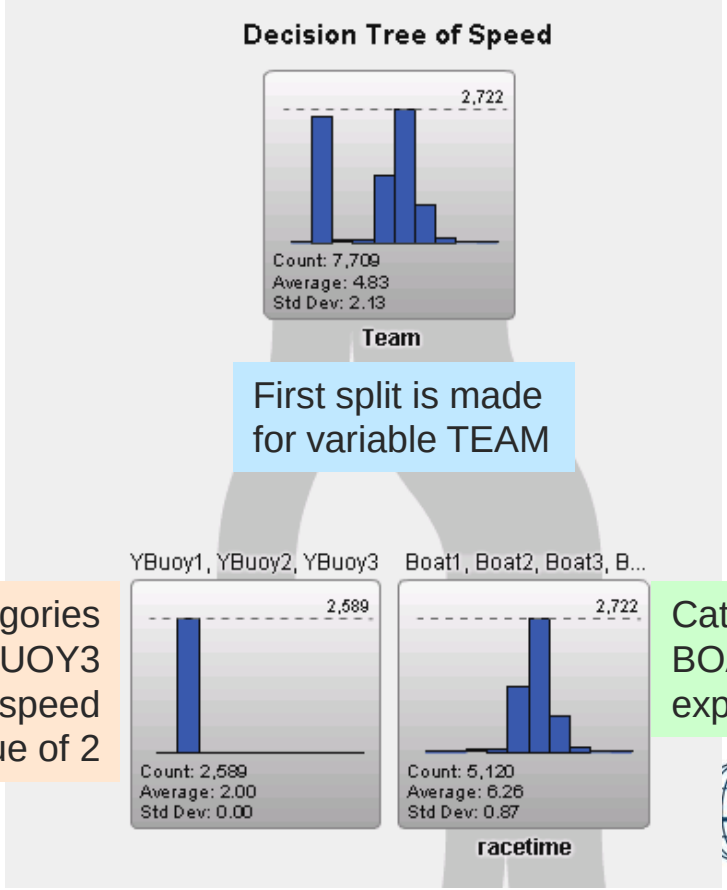
Distribution of the speed in knots

- Distribution of speed between 4 and 9 knots, makes sense for this type of sailboat
- Large accumulation of data points at 2 knots!



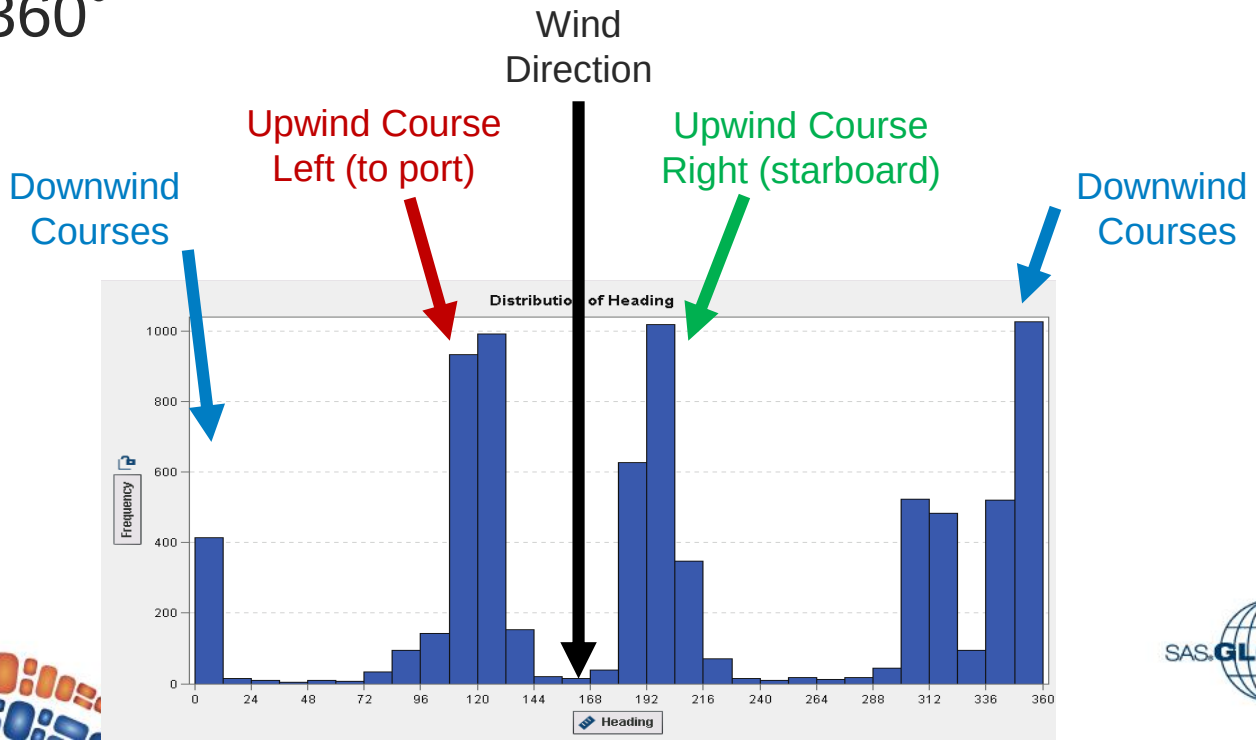
Using a decision tree to find the reason

1. Use variable Speed_in_Knots as a target variable
2. All other variables as input variables
3. The decision tree automatically segments by the most influential variable(s)



Using interactive data analysis to uncover impossible value combinations

- Distribution of compass heading looks fine: values from 0° to 360°



Business insight with interactive data analysis

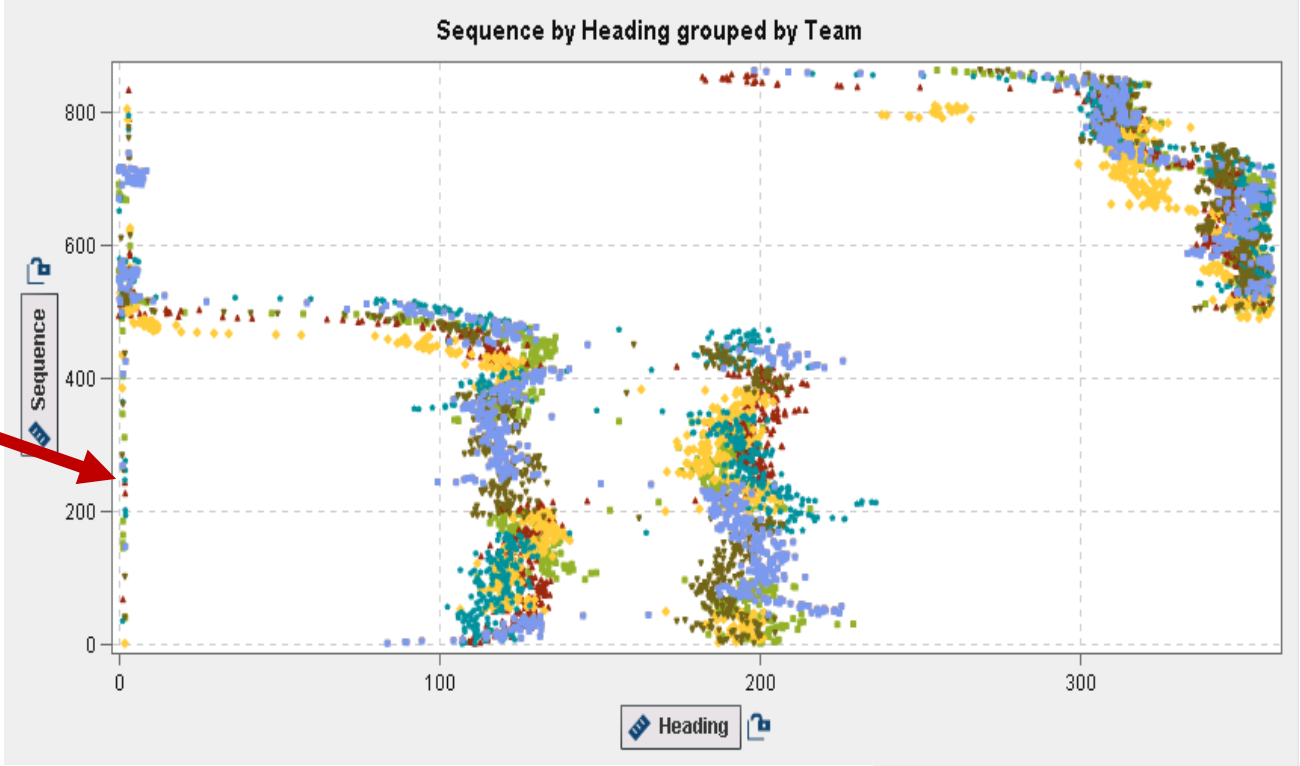
Compass Headings around 2° are selected

Surprising to see single track points with heading 2° also in the upwind section

While sailing upwind to 200°, boats cannot turn northward for 2 seconds!

The sequence plot reveals the same picture

Outliers occur for all teams throughout the upwind course!



Downwind Course

Upwind Course

Note: These are outliers from a *business perspective*. Technically the values between 0° and 360° are fine.

Drilling to the source data reveals the reason

- Drilling to the source data reveals the reason:
  - Compass Headings with two zeros (.00) after the decimal point are output as integer values
  - The data integration program that reads this data into a SAS data set did not consider such a situation
    - » Integer values are shifted to 2 decimal points. 198.00 → 1.98

```
"2009-05-21T14:04:40+02:00" heading="199.16" speed="5.9
"2009-05-21T14:04:42+02:00" heading="197.26" speed="5.9
"2009-05-21T14:04:44+02:00" heading="200.01" speed="5.7
"2009-05-21T14:04:46+02:00" heading="200.18" speed="5.7
"2009-05-21T14:04:48+02:00" heading="205.77" speed="5.5
"2009-05-21T14:04:50+02:00" heading="198" speed="5.405"
"2009-05-21T14:04:52+02:00" heading="205.26" speed="5.6
"2009-05-21T14:04:54+02:00" heading="195.28" speed="5.5
"2009-05-21T14:04:56+02:00" heading="198.07" speed="5.5
"2009-05-21T14:04:58+02:00" heading="204.78" speed="5.5
```

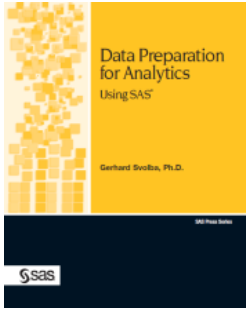
Summary

- Differentiate between *regular data quality* and *data quality for analytics*!
- Analytic methods have additional requirements on data quality - they also offer methods to profile and improve data quality
- SAS macros and SAS sample programs help you to profile your data in a very powerful way
- SAS® Visual Analytics offers powerful methods to interactively profile your data

Contact Information

Gerhard Svolba
Analytic Solution Architect
SAS-Austria
sastools.by.gerhard@gmx.net
http://www.sascommunity.org/wiki/Gerhard_Svolba
LinkedIn – XING – PictureBlog

Data Quality for Analytics Using SAS
SAS Press 2012
http://www.sascommunity.org/wiki/Data_Quality_for_Analytics



Data Preparation for Analytics Using SAS
SAS Press 2006
http://www.sascommunity.org/wiki/Data_Preparation_for_Analytics

My Favorite Business Case Studies With SAS Analytics
SAS Press, expected 2016/2017

SAS GLOBAL FORUM

Session ID SAS1440

SAS® GLOBAL FORUM

April 26-29
Dallas, TX